SPARSE DECOMPOSITION OF AUDIO SIGNALS USING A PERCEPTUAL MEASURE OF DISTORTION. APPLICATION TO LOSSY AUDIO CODING.

Ichrak Toumi, Olivier Derrien

SPARSE DECOMPOSITION OF AUDIO SIGNALS USING A PERCEPTUAL MEASURE OF DISTORTION. APPLICATION TO LOSSY AUDIO CODING

Ichrak Toumi, *

CNRS LMA
31 chemin Joseph-Aiguier
13402 Marseille Cedex 20
toumi@lma.cnrs-mrs.fr

Olivier Derrien,

Université de Toulon & CNRS LMA
31 chemin Joseph-Aiguier
13402 Marseille Cedex 20
derrien@lma.cnrs-mrs.fr

ABSTRACT

State-of the art audio codecs use time-frequency transforms derived from cosine bases, followed by a quantification stage. The quantization steps are set according to perceptual considerations. In the last decade, several studies applied adaptive sparse time-frequency transforms to audio coding, e.g. on unions of cosine bases using a Matching-Pursuit-derived algorithm [1]. This was shown to significantly improve the coding efficiency. We propose another approach based on a variational algorithm, i.e. the optimization of a cost function taking into account both a perceptual distortion measure derived form a hearing model and a sparsity constraint, which favors the coding efficiency. In this early version, we show that, using a coding scheme without perceptual control of quantization, our method outperforms a codec from the literature with the same quantization scheme [1]. In future work, a more sophisticated quantization scheme would probably allow our method to challenge standard codecs e.g. AAC.

Index Terms– *Audio coding, Sparse approximation, Iterative thresholding algorithm, Perceptual model.*

1. INTRODUCTION

Actually, state-of-the-art lossy audio coders (e.g. MP3, AAC, OGG) globally share the same structure [2]: The signal is first processed using a time-frequency (TF) transform, the transform coefficients are quantized according to a psycho-acoustic model and finally lossless binary coded. The TF transform is usually adaptive (i.e. it can switch between two pre-defined resolutions), perfectly invertible and introduces no redundancy. The most popular choice is the Modifed Discrete Cosine Transform (MDCT) [3]. Optimizing the quantization is performed in the frequency domain (i.e. independently in each time slot corresponding to one analysis/synthesis long window or a block of consecutive short windows). This process takes into account a perceptual weighting of frequency components computed by a psychoacoustic model.

However, this structure introduces two main limitations: 1) the invertible and non-redundant (i.e. orthogonal) TF transform takes advantage of an aliasing-cancelation property [3]. But when some spectral components are removed, because they are considered perceptually less relevant that others, some disturbing artifacts may become audible (called *birdies*). This occurs usually at low bitrate, where many components have to be removed. 2) Optimizing in the frequency domain does not allow to take into account the time-domain properties of audition (e.g. the temporal masking effect), which implies sub-optimal performance. Some propositions have been made to compensate for these drawbacks (e.g. [4, 5]), but it only resulted in marginal improvements.

The most important breakthrough in the last few years consists in performing a sparse approximation of the signal on an over-complete dictionary of waveforms instead of using an orthogonal transform [1]. This does not guarantee perfect reconstruction, but this property is not mandatory in a lossy codec. This approach was proved to significantly improve the audio quality on some audio files, especially at low bitrate. The proposed method relies on a Matching-Pursuit (MP) derived algorithm, which was modified to reduce artifacts called *pre-echo*. In its basic version, the MP is optimal with respect to the minimum Mean Square Error (MSE) criterion. Some improvements have been proposed in order to introduce perceptual weights in MP, but this was not selected in [1] because it significantly increases the complexity. Recently, a more sophisticated method for sparse audio signal decomposition was proposed [6]. It uses a *variational* algorithm and includes a perceptual weighting. However, the optimization is still performed only in the frequency domain and its implementation remains uncomplete (no quantization and binary coding were proposed).

In this paper, we describe a new method that is partially inspired by [1] and [6]. Our algorithm performs a sparse approximation on a over-complete dictionary, more precisely a union of MDCT bases, as in [1]. The optimization uses a variational algorithm, which appears to be more flexible than MP with respect to the distortion constraint. Our main contribution is that this algorithm performs the optimization in the TF plane, using a new TF audition model based on recent studies [7]. Then, we apply a quantization and binary coding scheme from the literature to evaluate the potential of this method in an audio-coding context. In [1], two codecs are described: #1 uses a simple *bit-plane* algorithm as a quantization and binary coding stage and #2 uses a so-called *Perceptual bit-plane* algorithm. It was shown that codec #1 usually does not perform as well as AAC, but codec #2 is able to outperform AAC, especially at low bitrate. In this paper, we only implement the simple bit-plane algorithm, because using the perceptual version is not straightforward in our coding scheme. For the same bitrates, we compare the audio quality for our codec and codec #1. In further works, we plan to improve our method by implementing a perceptual quantizer.

This paper is organized as follows: In section 2, we present the general method for sparse decomposition. In section 3, we motivate and describe our implementation. And in section 4, we compare our method to codec #1 described in [1].

* This work was supported by the joint French ANR and Austrian FWF project "POTION", refs. ANR-13-IS03-0004-01 and FWF-I-1362-N30.

2. THE METHOD FOR SPARSE DECOMPOSITION

Let's consider a block of audio samples noted as a vector $\mathbf{x}$ of size $1 \times N$, N being the number of samples. $\mathbf{x}$ can indifferently represent the whole signal, or a time-segment. We define the coding dictionary as a matrix $\mathbf{S}$ of size $M \times N$. Each row can be interpreted an elementary waveform, often called *atom*. The reconstructed signal is a linear combination of atoms, which can be written as:

$$\hat{\mathbf{x}} = \mathbf{a}\,\mathbf{S} \tag{1}$$

$\hat{\mathbf{x}}$ is the reconstructed signal and $\mathbf{a}$ is a vector of coefficients of size $1 \times M$, M being the number of atoms in the dictionary. This general scheme applies to state-of-the-art audio codecs: $\mathbf{S}$ is determined by the TF transform and $\mathbf{a}$ represents the transform coefficients to be quantized and coded. If we assume an invertible dictionary, $M = N$ and the coefficients can be easily computed using:

$$\mathbf{a} = \mathbf{x}\,\mathbf{S}^{-1}$$

This is the case for instance with TF transforms derived from the Fourier transform (DFT, DCT). An over-complete dictionary implies that $M > N$, which means that $\mathbf{S}$ is not invertible. In the general case, there is no $\mathbf{a}$ such that $\hat{\mathbf{x}} = \mathbf{x}$. Then, $\mathbf{a}$ must be computed using an iterative algorithm, with respect to a pre-defined distortion measure. For instance, the MP algorithm [8] finds $\mathbf{a}$ which minimizes the mean-square-error (MSE):

$$D(\mathbf{a}) = \|\mathbf{a}\,\mathbf{S} - \mathbf{x}\|^2$$

For audio signals, MSE is known to be poorly correlated to the auditory perception of distortion. In [6], a more efficient measure is proposed based on the concept of perceptual weights for each frequency component.

Here, we propose a more general formulation of the perceptual weighting technique. This comes from the observation that reasonably good perceptual distortion measures have already been proposed, e.g. the mean Noise-to-Mask Ratio (NMR) [9]. But these measures are usually associated to an analysis filterbank that follows the characteristics of human perception. Thus, we introduce a second matrix $\mathbf{P}$ of size $N \times K$ that represents a perceptual transform. We assume that $K > N$, i.e. this transform is redundant. In the general case, $\mathbf{P}$ and $\mathbf{S}$ are not directly related, because a good perceptual filterbank (with respect to the accuracy of perceptual modeling) is generally not so good for coding purposes (with respect to the efficiency of coding). The coefficients of this transform corresponding to the signal $\mathbf{x}$ can be written as:

$$\mathbf{p}_x = \mathbf{x}\,\mathbf{P}$$

We define the perceptual distortion measure as a weighted version of MSE computed in the perceptual-transform domain:

$$D_p(\mathbf{a}) = \|\,(\mathbf{p}_{\hat{x}} - \mathbf{p}_x)\,\mathbf{\Delta}_x\|^2 = \|\,(\mathbf{a}\mathbf{S} - \mathbf{x})\,\mathbf{P}\mathbf{\Delta}_x\|^2$$

Where $\mathbf{\Delta}_x = \mathrm{diag}\,(\mu_k(\mathbf{x}))$ is a diagonal matrix of size $K \times K$ containing perceptual weights μ_k associated to the components of $\mathbf{p}_x$. These weights depend on $\mathbf{x}$ and can be computed using a hearing model of masking. If we note $T_k(\mathbf{x})$ the masking threshold values corresponding to the signal $\mathbf{x}$, we choose $\mu_k(\mathbf{x}) = (T_k(\mathbf{x}))^{-\frac{1}{2}}$. Then, the perceptual distortion $D_p(\mathbf{a})$ can be interpreted as a mean NMR. Note that, if the perceptual transform performs a TF analysis with a sufficiently good time precision, both temporal and frequential masking can be modeled.

The general coding problem then consists in finding, for any input signal $\mathbf{x}$, the optimal vector of coefficients $\mathbf{a}$ which minimizes the perceptual distortion measure $D_p(\mathbf{a})$. The existence of a solution depends on the matrix $\mathbf{K} = \mathbf{SP}$ called the *mixture matrix*. It is quite easy to prove that the maximum value for the rank of $\mathbf{K}$ is N. The minimization problem would have a single minimum if $\mathbf{K}$ would have a full-rank (i.e. $\min\{M, K\}$), which is never the case here because $N < \min\{M, K\}$. In other words, there are many equivalent solutions. We must add a sparsity constraint on $\mathbf{a}$ to select the best solution, i.e. we add a regularization term Ψ to the distortion measure and then we minimize the following objective function:

$$\Phi(\mathbf{a}) = D_p(\mathbf{a}) + \Psi(\mathbf{a})$$

This class of problem can be solved using an iterative thresholding algorithm [10, 11]. However, when $\mathbf{K}$ has not a maximum rank (i.e. N), the convergence is not always satisfactory.

In the literature, Ψ is often assimilated to an ℓ_α norm on the coefficients, where α is usually equal to 1. Although ℓ_1 norm does not directly measure the amount of zero coefficients, one usually assume that the minimum ℓ_1 norm corresponds to a sparse solution [10]. Then,

$$\Phi(\mathbf{a}) = \|\,(\mathbf{a}\mathbf{S} - \mathbf{x})\,\mathbf{P}\mathbf{\Delta}_x\|^2 + \lambda\|\mathbf{a}\|_\alpha \tag{2}$$

where the λ parameter allows to set the tradeoff between the distortion constraint and the sparsity constraint.

3. OUR PROPOSED IMPLEMENTATION

Figure 1 shows a block-diagram of our coder. It is mainly composed of 3 parts: The adaptive analysis over a coding dictionary which is a union of MDCT bases, a psycho-acoustic model and a quantization and coding section. The main point consists in finding a suitable coding dictionary $\mathbf{S}$ and a perceptual transform $\mathbf{P}$ such that $\mathbf{K} = \mathbf{SP}$ has a maximum rank.

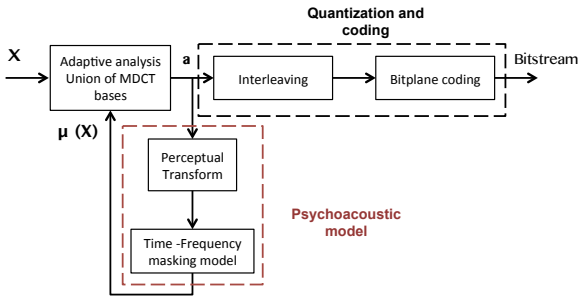

Figure 1: *Diagram of the proposed coder.*

3.1. Union of MDCT bases

First, we assume a coding scheme that splits the audio file in time-segments of reasonably long duration, called *macro blocks*, in order to be compatible with audio streaming applications (e.g. over the internet). Thus, we assume that the signal $\mathbf{x}$ corresponds to a macro-block. This approach requires an overlap between macro-blocks, and we wish that it does not introduce redundancy. Thus, a suitable choice for $\mathbf{S}$ is a union of MDCTs of different sizes, in the

same way as in [1]. In this paper, the authors proposed a union of 8 MDCTs: 64, 128, 256, 512, 1024, 2048, 4096 and 8192 bands. In this study, we restrained our dictionary to a union of 2 MDCTs with medium numbers of bands: 256 and 2048. In this implementation, one macro-bloc is composed of 16 short-windows (for 256-band) and 2 long-windows (for 2048-band). Choosing the number of MDCT sizes is tricky: More MDCTs, especially longer ones, would increase the efficiency of the decomposition, but would also degrade the efficiency of the coding stage: More MDCTs means more coefficients in $\mathbf{a}$, and even with the same number of non-zero coefficients, the additional zero coefficients require more coding bits.

The vector of MDCT coefficients $\mathbf{a}$ is made of two parts:

$$\mathbf{a} = [\mathbf{a_{256}}\ \mathbf{a_{2048}}]$$

where $\mathbf{a_{256}}$ stands for the coefficients of the short MDCT and $\mathbf{a_{2048}}$ for the long MDCT. The time-support corresponding to one macro-block is determined by the long MDCT, i.e. $N_{\text{mb}} = 3 \times 2048 = 6144$ samples. But practically, the optimization must be performed on an analysis-segment longer than a macro-block, in order to avoid sharp variations of coding parameters between successive macro-blocks. Thus, N corresponds to the length of one analysis segment. Here, we chose $N = 7 \times 2048 = 14336$ samples. This is illustrated on figure 2. Both $\mathbf{a_{2048}}$ and $\mathbf{a_{256}}$ are of size $1 \times M_{\text{as}}$, with $M_{\text{as}} = 6 \times 2048 = 12288$, and $\mathbf{a}$ is of size $1 \times 2M_{\text{as}}$, which means $M = 2M_{\text{as}} = 24576$. But practically, only 8192 coefficients (in the middle) are actually coded. The first and last 8192 coefficients are discarded. This part of signal will be analyzed and coded respectively in the previous and the next macro-blocks.

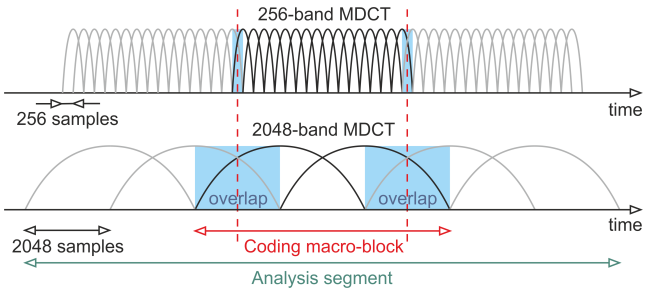

Figure 2: *MDCT windows in the coding dictionary for one macro-block, and corresponding analysis segment.*

3.2. Perceptual model

For the perceptual transform, we seek for a filterbank that follows the audition, but also with nice mathematical properties. The rank constraint on $\mathbf{K}$ requires that $\mathbf{P}$ is a full-rank matrix, i.e. its rank is N. A good choice is the ERB-MDCT, which is a near-orthogonal transform that follows the ERB frequency-scale [12]. In this transform, the time-resolution is adapted to the frequency resolution in each frequency band: Low frequencies have a high frequency-resolution and a low time-resolution, whereas high frequencies have a lower-frequency resolution and a higher time-resolution. We choose an ERB-MDCT with one band per ERB, which corresponds to $K = 15126$. One can check that we actually have $K > N$.

The square-value of ERB-MDCT coefficients is interpreted as the temporal and spectral density of energy. To obtain a masking threshold, we convolve the TF energy image with the TF masking kernel described in [7], and plotted on figure 3. We assume the additivity of masking patterns in the energy scale, which is known to sometimes under-estimate the masking level. Finally, we keep the minimum value of the masking threshold and the absolute threshold of hearing in quiet as described in MPEG #1 psycho-acoustic model [13], and get an estimation of the $\mu_k(\mathbf{x})$.

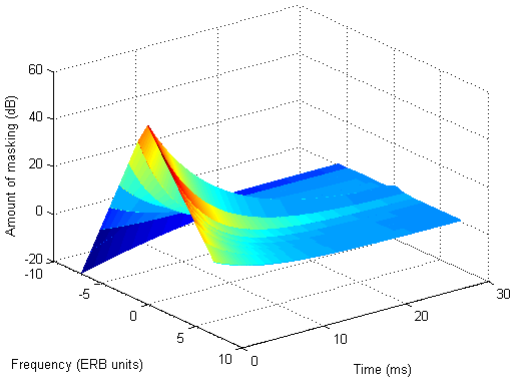

Figure 3: *Time-frequency masking kernel.*

3.3. Adaptive decomposition

For the optimization in the adaptive analysis step, we used the FISTA algorithm proposed by Beck et al. in [11]. The algorithm includes a gradient step followed by a shrinkage/soft-thresholding step. The general update rule in FISTA is

$$\mathbf{a}_i = \mathrm{T}_{\lambda\gamma}\left(G(\mathbf{y}_i)\right)$$

where i is the iteration index, γ is the gradient stepsize, $G(.)$ is the gradient of the distortion term D_p, $\mathbf{y}_i$ is a specific linear combination of the previous two iterations $\{\mathbf{a}_{i-1}, \mathbf{a}_{i-2}\}$ and $\mathrm{T}_{\lambda\gamma} : \mathbb{R}^M \rightarrow \mathbb{R}^M$ is the shrinkage operator:

$$\mathrm{T}_{\lambda\gamma}(\mathbf{a})_m = \mathrm{sign}(a_m)\ \max\left(0, |a_m| - \lambda\gamma\right)$$

A typical condition which guarantees the convergence to a single minimiser $\mathbf{a}^\star$ is $\gamma \in \left[0, 1/\left\|\mathbf{K}^T\mathbf{K}\mathbf{\Delta}_x\right\|\right]$.

In the following, we assume for each macro-block:

- $\alpha = 1$, which amounts to perform a soft thresholding on the coefficients $\mathbf{a}$.
- $\gamma = 1/\left\|\mathbf{K}^T\mathbf{K}\mathbf{\Delta}_x\right\|$.
- For $i = 0$, $\mathbf{a}_0 = 0_{1\times M}$.

A crucial parameter for the optimization process is λ. We found out that a constant value for λ does not always produce the same sparsity ratio for $\mathbf{a}$. Thus, the suitable value for λ has to be set for each macro-block in order to get a constant sparsity ratio. Our goal was to get approximately the same rate of non-zero coefficients in $\mathbf{a}$ as with quantized MDCT coefficients in a MPEG-AAC bitstream. We measured that, for a monophonic AAC operating at 48 kbps, approximately 30 % of the MDCT coefficients

are equal to zero. Then, we target a sparsity ratio of 15 %, since our dictionary has a redundancy factor of 2.

Furthermore, we found out that setting the same value of λ for $\mathbf{a_{256}}$ and $\mathbf{a_{2048}}$ was not the optimal choice. Thus, we rewrite equation (2) as

$$\Phi(\mathbf{a}) = \| \left(\mathbf{aS} - \mathbf{x}\right) \mathbf{P} \mathbf{\Delta}_x \|^2 + \lambda_1 \|\mathbf{a_{256}}\|_\alpha + \lambda_2 \|\mathbf{a_{2048}}\|_\alpha \quad (3)$$

where λ_1 and λ_2 are respectively the regularization parameters corresponding to $\mathbf{a_{256}}$ and $\mathbf{a_{2048}}$. From informal listening tests, it appears that the best audio quality is obtained with $\lambda_1 > \lambda_2$, which means more non-zero coefficients in the long MDCT than in the short one.

In the literature, the choice of a suitable value for λ is known to be a difficult problem [14, 15], that would deserve a more specific study by itself. In this work, we use a simplified formula from [10] which is valid in the following conditions: (i) $\mathbf{K}$ is the identity operator and (ii) $\alpha = 1$:

$$\mathrm{T}_\lambda(\mathbf{x}) = \begin{cases} \mathbf{x} + \frac{\lambda}{2} & \text{if} \quad \mathbf{x} \leq -\frac{\lambda}{2}, \\ 0 & \text{if} \quad |\mathbf{x}| < \frac{\lambda}{2}, \\ \mathbf{x} - \frac{\lambda}{2} & \text{if} \quad \mathbf{x} \geq \frac{\lambda}{2}. \end{cases} \quad (4)$$

Then, assuming that $\forall k \in \{1 \cdots K\}$, $|\mathbf{p}_x(k)\mathbf{\Delta}_x(k)| \geq \frac{\lambda}{2}$, we get:

$$\lambda = 2Y_s \left\lfloor K(1 - sp) + 1 \right\rfloor ;$$

where sp is the amount of sparsity, Y_s are the values of $|\mathbf{p}_x \mathbf{\Delta}_x|$ sorted in increasing order, and $\lfloor . \rfloor$ denotes the rounding operator towards $-\infty$. Although $\mathbf{K}$ is not the identity operator in our case, this formula still gives a good approximation of the optimal λ in each macro-block. Thus, controlling sparsity only requires to a adjust the parameter sp.

3.4. Coefficients quantization and coding

When the coefficients $\mathbf{a}$ are computed, we carry out the quantification and binary coding preceded by an interleaving step which aims to group together the coefficients that are close in the time-frequency plane, and to make the coding algorithm more efficient. We chose the same binary coder as in [1]: a bit-plane coder, which is a special case of adaptive Rice-Golomb codes. It also performs an implicit quantization: the number of coding bits corresponds to the number of bit-planes used for coding. This method is particularly efficient when the sequence to be encoded exhibits long ranges of zeros, and thus is well suited for encoding sparse vectors. However, our interleaving scheme, as illustrated on figure 4, is different from the one described in [1]. The idea is to group the low frequencies at the beginning and the high frequencies at the end. We found out that our method reduces the bitrate, because it favors long ranges of zeros at the end.

4. RESULTS AND DISCUSSION

To test our method, we used the same audio material as in [1]. This collection of 4 musical pieces, sampled at 44.1 KHz, is described on table 1. Note that all the audio signals mentioned in this paper (unprocessed, re-synthesized and coded/decoded signals) can be downloaded from the companion webpage of this paper: http://potion.cnrs-mrs.fr/dafx2015.html.

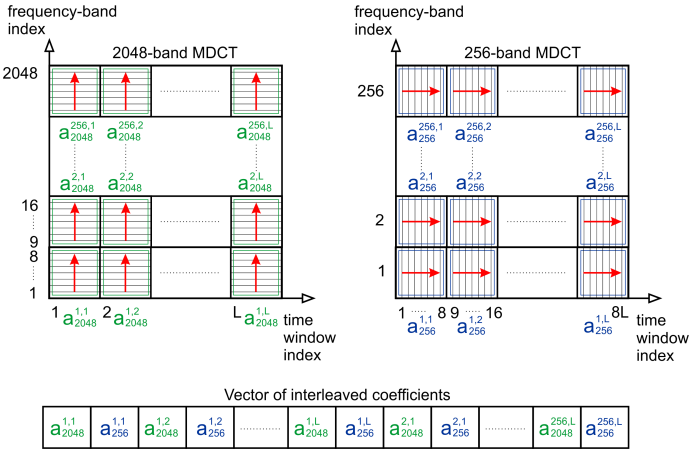

Figure 4: *Interleaving scheme for MDCT coefficients.*

Test signal	Description	Duration (s)
harp	Harpsichord	8.0
bagp	Bagpipes	11.1
orch	Orchestral piece	12.7
popm	Contemporay pop music	11.6

Table 1: *Test signals used for evaluation.*

4.1. Performance of the adaptive decomposition algorithm

In a first step, we analyze the results of the adaptive decomposition algorithm on the *harp* signal. For evaluation purpose, we re-synthesize the audio signal without quantization by applying equation (1). As explained in section 3.3, the sparsity ratio is about 10% on 256-band MDCT coefficients and 20% on 2048-band MDCT coefficients.

We can see on the spectrograms plotted on figure 5 that the original and re-synthesized signals are very similar. However, the reconstructed signal has less energy on the regions of the TF plane that are between the partials, which should correspond to masked regions. One can also notice that partials sometimes cross vertical structures corresponding to attacks, which can be associated to *pre-echo*. This is most probably due to the fact that our short MDCT has 256 bands, whereas in [1], the shortest MDCT has only 64 bands.

We also plot the *TF maps* of the coefficients $\mathbf{a_{256}}$ and $\mathbf{a_{2048}}$ on figure 6. In other words, we plot the magnitude of MDCT coefficients in the TF plane for both MDCT bases separately. First, one can see that the long MDCT exhibits a good frequency resolution, but a poor time-resolution. The short MDCT has opposite properties. One can also see that partials are essentially represented by long MDCT coefficients, and attacks by short MDCT coefficients. This proves that our optimization algorithm dispatches the energy between MDCT bases in an accurate way. However, this process is not perfect: some partials are represented in the short MDCT.

Finally, we plot on figure 7 the spectrograms of the re-synthesized signal after quantization and coding at 48 kbps and 24 kbps, for the same original test signal. At 48 kbps, one can see that the spectrogram of the re-synthesized signal is highly similar to the spectrogram before quantization and coding. However, at low

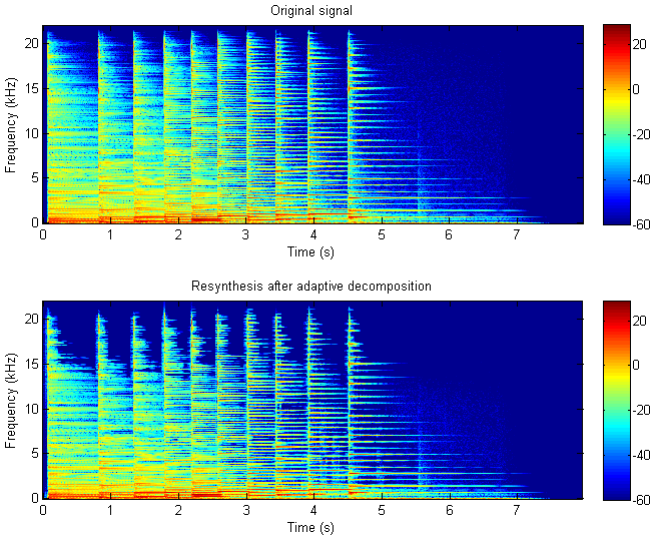

Figure 5: *Spectrograms of the* harp *signal*

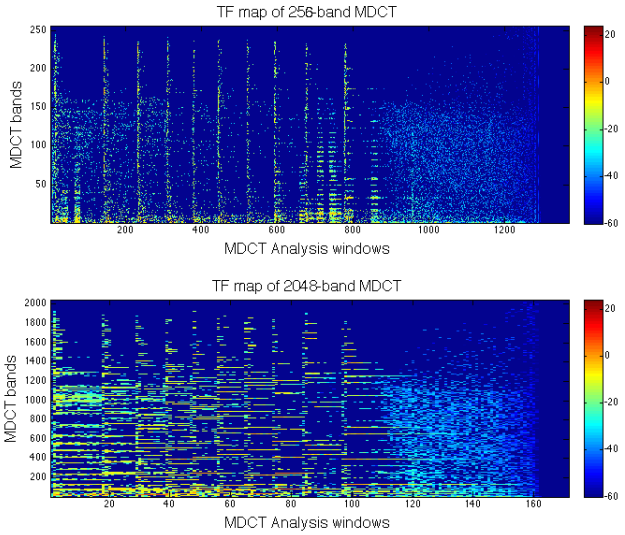

Figure 6: *TF maps of coefficients for the* harp *signal. Up: short MDCT. Down: long MDCT.*

bitrate (24 kbps), the spectrogram is somewhat different, with a slight degradation in high frequencies and less energy between partials.

4.2. Evaluation of audio quality

In order to evaluate the impact of our method on the audio quality and to compare our codec (denoted in section *codec A*) to the codec #1 proposed in [1], we organized a MUSHRA listening test [16] on the four audio signals described in table 1, with scores ranging from 0 (very bad quality) to 100 (excellent quality). Six versions of the test signals were evaluated by a total number of 16 listeners:

- A hidden reference,
- A 4 kHz low-pass anchor,
- Two coded versions with codec A at 24 and 48 kbps,
- Two coded versions with codec #1 at 24 and 48 kbps.

The listeners were post-screened according to the scoring of the reference: If the score attributed to the reference is lower than 80, the listener was judged unreliable and discarded. Finally, only 10 subjects were kept.

The results of MUSHRA listening tests are plotted on figure 8 and 9, respectively at 24 and 48 kbps (mean values and 95% confidence intervals). For both bitrates, one can see that results highly depend on the test signal. Signals with naturally sparse spectrums (*harp*, *bagp*) are associated to high audio quality at both bitrates. This can be interpreted as follows: The amount of perceptually relevant information in these signals is relatively low, and a good reconstruction in achievable at low bitrate. For *bagp*, the results obtained with both codecs are similar at 24 and 48 kbps but for *harp*, codec #1 is better rated. According to the listeners' feedback, this was mainly due to the fact that attacks are not as sharp with our codec as with codec #1, which is probably related to the length of the short MDCT. Signals with more dense spectrums (*orch*, *popm*) are associated to lower audio quality, especially a 24 kbps, probably because the amount of perceptually relevant information in these signals is higher. At 24 kbps, both codecs perform similarly on these two signals, but at 48 kbps, our codec is slightly better on *popm*, and much better on *orch*. Note that, at 24 kbps, both codecs were rated lower than the anchor, because a reduced band-pass was judged less disturbing than a high level of coding artifacts.

Finally, on average over the 4 test signals, our codec was rated slightly worse at 24 kbps, only because it does not perform as well on *harp*, and rated significantly better at 48 kbps, because it performs better on *orch* and *popm*.

5. CONCLUSION

In this paper, we described a method that performs an adaptive decomposition of audio signal on redundant coding dictionary (a union of 2 MDCT bases) using a variational algorithm, i.e. the optimization of a cost function taking into account both a perceptual distortion measure derived form a hearing model and a sparsity constraint. We applied a simple quantization and coding scheme from the literature (simple bit-plane coder) and compared the final audio quality to one achieved by codec #1 in [1], which uses the same quantization and coding scheme. We show that at low bitrate (24 kbps), our codec performs worse on signal with many sharp attacks, and performs similarly on other signals. This can be explained by the fact that our short MDCT has only 256 bands. At medium bitrate (48 kbps), our codec performs better on audio signals with a dense spectrum. This point is particularly interesting since in [1], codec #2 was able to outperform AAC but not on signals with dense spectrum (*orch* and *popm*), despite the use of a complex perceptual quantization scheme. Thus, these results are promising: As it is, our codec can not outperform codec #2 or AAC, but it may seriously challenge both if we design a suitable perceptual quantization scheme. This would also require to extend the coding dictionary by adding a shorter MDCT (128 or 64 bands).

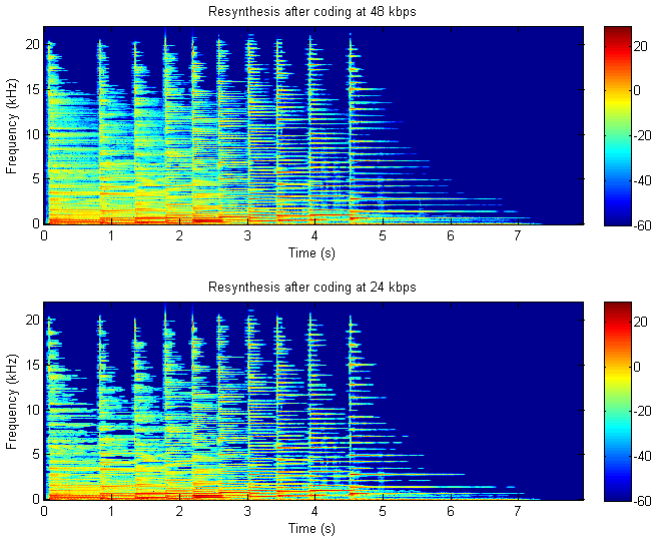

Figure 7: *Spectrograms of the* harp *signal after quantization and coding.*

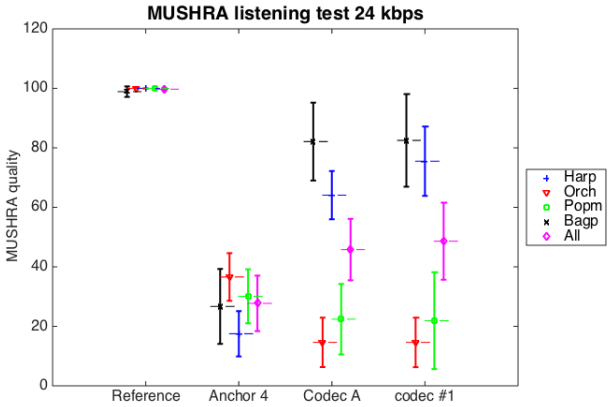

Figure 8: *Results of MUSHRA listening test at 24 kbps.*

6. REFERENCES

[1] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Tr. ASLP*, vol. 16, no. 8, pp. 1361–1372, Nov. 2008.

[2] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, Wiley, 2007.

[3] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Tr. ASSP*, vol. 34, no. 5, pp. 1153–1161, Oct. 1986.

[4] O. Derrien and L. Daudet, "Reduction of artefacts in mpeg-aac with mdct spectrum regularisation," in *Proc. 116th AES Convention*, Berlin, Germany, May 08-11, 2004.

[5] H. Najaf-Zadeh, H. Lahdidli, M. Lavoie, and L. Thibault, "Use of auditory temporal masking in the mpeg psychoacoustic model 2," in *Proc. 114th AES Convention*, Amsterdam, The Netherlands, March 22-25, 2003.

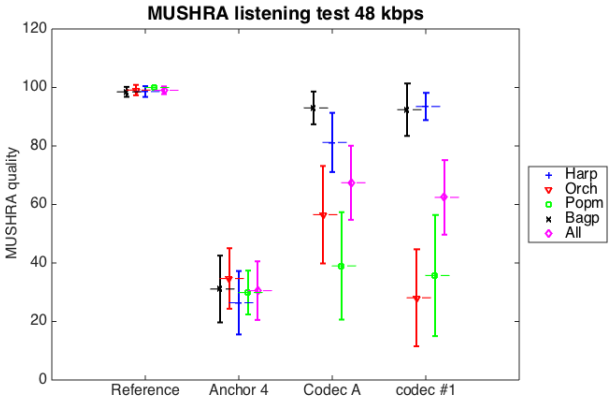

Figure 9: *Results of MUSHRA listening test at 48 kbps.*

[6] M.G. Christensen and B.L. Sturm, "A perceptually reweighted mixed-norm method for spaser approximation of audio signals," in *Proc. ASILOMAR*, Pacific Grove, California, Nov. 06-09, 2011.

[7] T. Necciari, *Time-Frequency Masking: Psychoacoustical Measures and Application to the Analysis-Synthesis of Sound Signals*, Ph.D. thesis, University of Aix-Marseille (in french), 2010.

[8] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Tr. SP*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] K. Brandenburg and T. Sporer, "NMR and Masking Flag : Evaluation of quality using perceptual criteria," in *Proc. 11th Int. Conf. of the AES*, 1992.

[10] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[11] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[12] O. Derrien, T. Necciari, and P. Balazs, "A quasi-orthogonal, invertible and perceptually relevant time-frequency transform for audio coding," in *Proc. EUSIPCO*, Nice, France, Aug. 31 - Sept. 04, 2015.

[13] International Organization for Standardization, *ISO/IEC 11172-3 (Information Technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s) - Part 3: Audio*, 1993.

[14] P. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, Society for Industrial and Applied Mathematics, 1998.

[15] Robert Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[16] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality levels of coding systems," Tech. Rep., International Telecommunication Union, Geneva, Switzerland, 2003.